Empirische Varianz und Stichprobenvarianz

Sei x_i eine Merkmalsausprägung von $i = 1...n$ Merkmalsausprägungen, dann ist die Varianz definiert als die Summe der quadrierten Abweichungen von ihrem Mittelwert, geteilt durch die Anzahl der Merkmalsausprägungen:

(1) $$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

Dabei ist der Mittelwert, die durchschnittliche Merkmalsausprägung:

(2) $$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Eine Varianz, in die alle Elemente der Grundgesamtheit einfließen, sei als empirische Varianz bezeichnet. Beschränkt sich die statistische Erhebung dagegen nur auf einen Teil der Grundgesamtheit, ist die Varianz eine Stichprobenvarianz.

In einer empirischen Varianz wird die durchschnittliche Merkmalsausprägung aus *allen* Merkmalsausprägungen der Grundgesamtheit ermittelt. Eben deswegen ist dieser Wert der richtig ermittelte Durchschnitt, die wahre durchschnittliche Merkmalsausprägung.

Bei einer Stichprobenvarianz aber wird der Durchschnittswert aus den Merkmalsausprägungen der Erhebungseinheiten ermittelt, und das sind bei einer Stichprobe nicht alle Elemente der Grundgesamtheit. Die wahre durchschnittliche Merkmalsausprägung bleibt bei einer Stichprobe unbekannt. Um die richtige Varianz einer Stichprobe nach Gleichung (1) auszurechnen, benötigt man aber den wahren Mittelwert der Merkmalsausprägungen. Wird dieser als $\overline{x_w}$ bezeichnet und die aufgrund dieses wahren Mittelwertes errechnete Stichprobenvarianz als s_p^2, so gilt:

(3) $$s_p^{\,2} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x_w} \right)^2$$

Da aber $\overline{x_w}$ unbekannt ist, lässt sich die wahre Stichprobenvarianz mit dieser Formel allerdings nicht berechnen. Man muss es mit $\bar{x}$ versuchen, der durchschnittlichen Merkmalsausprägung der Stichprobe.

Um eine Verbindung zwischen $\bar{x}$ und $\overline{x_w}$ herzustellen, wird der in Gleichung (3) enthaltene Ausdruck

$$\sum_{i=1}^{n} \left(x_i - \overline{x_w} \right)^2$$

gesondert betrachtet. In diesem Ausdruck wird $\bar{x}$ einerseits addiert und andererseits subtrahiert, sodass sich der Wert nicht ändert:

$$\begin{aligned} \sum_{i=1}^{n} \left(x_i - \overline{x_w} \right)^2 &= \sum_{i=1}^{n} \left(x_i - \bar{x} + \bar{x} - \overline{x_w} \right)^2 \\ &= \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right) + \left(\bar{x} - \overline{x_w} \right) \right]^2 \end{aligned}$$

Die Ausdruck in der eckigen Klammer wird ausmultipliziert:

$$\sum_{i=1}^{n} \left(x_i - \overline{x_w} \right)^2 = \sum_{i=1}^{n} \left[\left(x_i - \bar{x} \right)^2 + 2 \left(x_i - \bar{x} \right) \cdot \left(\bar{x} - \overline{x_w} \right) + \left(\bar{x} - \overline{x_w} \right)^2 \right]$$

Die Summanden werden voneinander getrennt:

(4) $$\sum_{i=1}^{n} \left(x_i - \overline{x_w} \right)^2 = \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 + 2 \sum_{i=1}^{n} \left(x_i - \bar{x} \right) \cdot \left(\bar{x} - \overline{x_w} \right) + \sum_{i=1}^{n} \left(\bar{x} - \overline{x_w} \right)^2$$

Der in Gleichung (4) enthaltene Ausdruck $\sum_{i=1}^{n}\left(x_i-\bar{x}\right)$ wird wiederum näher betrachtet:

(5) $$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\bar{x}$$

Hierin bedeutet der Ausdruck $\sum_{i=1}^{n}\bar{x}$, dass $\bar{x}$ für alle Werte von *i* jeweils einmal summiert werden muss, insgesamt also *n* Mal. Es gibt also *n* Summanden mit dem Wert $\bar{x}$, sodass

$$\sum_{i=1}^{n}\bar{x}=n\cdot\bar{x}$$

Dies in Gleichung (5) eingesetzt:

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=\sum_{i=1}^{n}x_i-n\cdot\bar{x}$$

Nach der Definition des Mittelwertes gemäß Gleichung (2) ist aber auch $\sum_{i=1}^{n}x_i=n\cdot\bar{x}$, sodass

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=n\cdot\bar{x}-n\cdot\bar{x}=0$$

Man erhält das zu erwartende Ergebnis, dass nämlich die Summe der Abweichungen vom Mittelwert gleich null ist. Damit ist in Gleichung (4) auch

$$2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(\bar{x}-\overline{x_w}\right)=0$$

Gleichung (4) reduziert sich also auf:

(6) $$\sum_{i=1}^{n}\left(x_i-\overline{x_w}\right)^2=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+\sum_{i=1}^{n}\left(\bar{x}-\overline{x_w}\right)^2$$

Für den in Gleichung (6) enthaltenen Ausdruck $\sum_{i=1}^{n}\left(\bar{x}-\overline{x_w}\right)^2$ gilt wieder, dass jeder Summand gleich groß ist, nämlich $\left(\bar{x}-\overline{x_w}\right)^2$, und dieser Summand für alle Werte von *i* addiert werden muss, insgesamt also *n* Mal. Somit ist

$$\sum_{i=1}^{n}\left(\bar{x}-\overline{x_w}\right)^2=n\cdot\left(\bar{x}-\overline{x_w}\right)^2$$

Dies in Gleichung (6) eingesetzt:

(7) $$\sum_{i=1}^{n}\left(x_i-\overline{x_w}\right)^2=\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\cdot\left(\bar{x}-\overline{x_w}\right)^2$$

Nunmehr kann Gleichung (7) in die ursprüngliche Gleichung (3) eingesetzt werden:

(8) $$s_p{}^2=\frac{1}{n}\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+\left(\bar{x}-\overline{x_w}\right)^2$$

Hierin Gleichung (1) eingesetzt:

(9) $$s_p^{\,2} = s^2 + \left(\overline{x} - \overline{x_w}\right)^2$$

Hierin Gleichung (2) eingesetzt:

$$s_p^{\,2} = s^2 + \left(\frac{1}{n}\cdot\sum_{i=1}^{n} x_i - \overline{x_w}\right)^2 = s^2 + \left(\frac{1}{n}\cdot\sum_{i=1}^{n} x_i - \frac{n}{n}\cdot\overline{x_w}\right)^2$$

Der Faktor $\frac{1}{n}$ wird ausgeklammert:

(10) $$s_p^{\,2} = s^2 + \frac{1}{n}\cdot\frac{1}{n}\cdot\left(\sum_{i=1}^{n} x_i - n\cdot\overline{x_w}\right)^2$$

Der in Gleichung (10) enthaltene Ausdruck $\left(\sum_{i=1}^{n} x_i - n\cdot\overline{x_w}\right)^2$ wird gesondert betrachtet und zunächst explizit formuliert:

$$\left(\sum_{i=1}^{n} x_i - n\cdot\overline{x_w}\right)^2 = \left(x_1 + x_2 + x_3 \ldots + x_n - n\cdot\overline{x_w}\right)^2$$

Da für die *n* Elemente x_i genau *n* Elemente $\overline{x_w}$ vorhanden sind, kann jedem x_i ein $\overline{x_w}$ zugeordnet werden:

$$\left(\sum_{i=1}^{n} x_i - n\cdot\overline{x_w}\right)^2 = \left[\left(x_1 - \overline{x_w}\right) + \left(x_2 - \overline{x_w}\right) + \left(x_3 - \overline{x_w}\right)\ldots + \left(x_n - \overline{x_w}\right)\right]^2$$

Die eckige Klammer wird ausmultipliziert:

$$\begin{aligned}\left(\sum_{i=1}^{n} x_i - n\cdot\overline{x_w}\right)^2 = {} & \left(x_1 - \overline{x_w}\right)^2 + \left(x_1 - \overline{x_w}\right)\cdot\left(x_2 - \overline{x_w}\right) + \left(x_1 - \overline{x_w}\right)\cdot\left(x_3 - \overline{x_w}\right)\ldots + \left(x_1 - \overline{x_w}\right)\cdot\left(x_n - \overline{x_w}\right)\\ & + \left(x_2 - \overline{x_w}\right)\cdot\left(x_1 - \overline{x_w}\right) + \left(x_2 - \overline{x_w}\right)^2 + \left(x_2 - \overline{x_w}\right)\cdot\left(x_3 - \overline{x_w}\right)\ldots + \left(x_2 - \overline{x_w}\right)\cdot\left(x_n - \overline{x_w}\right)\\ & + \left(x_3 - \overline{x_w}\right)\cdot\left(x_1 - \overline{x_w}\right) + \left(x_3 - \overline{x_w}\right)\cdot\left(x_2 - \overline{x_w}\right) + \left(x_3 - \overline{x_w}\right)^2 \ldots + \left(x_3 - \overline{x_w}\right)\cdot\left(x_n - \overline{x_w}\right)\\ & \vdots\\ & + \left(x_n - \overline{x_w}\right)\cdot\left(x_1 - \overline{x_w}\right) + \left(x_n - \overline{x_w}\right)\cdot\left(x_2 - \overline{x_w}\right) + \left(x_n - \overline{x_w}\right)\cdot\left(x_3 - \overline{x_w}\right) + \left(x_n - \overline{x_w}\right)^2\end{aligned}$$

Da jedes Element des zu quadrierenden Klammerausdrucks mit jedem Element der Summe multipliziert werden muss, wird jedes Element auch einmal mit sich selbst multipliziert und dadurch zum Quadrat erhoben. Die Summe dieser Quadrate ist also

$$\left(x_1 - \overline{x_w}\right)^2 + \left(x_2 - \overline{x_w}\right)^2 + \left(x_3 - \overline{x_w}\right)^2 \ldots + \left(x_n - \overline{x_w}\right)^2$$

Diese Summe ist wegen der Quadrierung in jedem Fall positiv.

Die übrigen Elemente bestehen aus den Produkten der Abweichungen jeweils verschiedener Werte von x_i. Diese Produkte können, je nachdem ob x_i über oder unter $\overline{x_w}$ liegt, teils positiv und teils negativ sein; das heißt, sie heben sich in der Summe mehr oder weniger auf. Die genaue Summe lässt sich nicht ermitteln, da $\overline{x_w}$ eben unbekannt ist. Wenn aber die x-Werte zufällig verteilt sind, kann man davon ausgehen, dass sich die positiven und negativen Abweichungen weitgehend aufheben,

sodass die Summe sich an null annähert, jedenfalls gegenüber den positiven Quadraten nicht weiter ins Gewicht fällt und somit vernachlässigt werden kann[1].

Wenn die x-Werte nicht zufällig verteilt sind, sollte man die gesamte Grundgesamtheit erheben oder herausfinden, welchem nicht-zufälligen funktionalen Zusammenhang die Werte unterliegen.

Hier wird die Voraussetzung zufällig verteilter x-Werte als gegeben erachtet und die Summe der nicht-quadratischen Produkte (der gemischten Klammern mit jeweils zwei verschiedenen x-Werten) als zu vernachlässigende Größe angesetzt, sodass

$$\left(\sum_{i=1}^{n} x_i - n \cdot \overline{x_w}\right)^2 = \left(x_1 - \overline{x_w}\right)^2 + \left(x_2 - \overline{x_w}\right)^2 + \left(x_3 - \overline{x_w}\right)^2 \ldots + \left(x_n - \overline{x_w}\right)^2 = \sum_{i=1}^{n}\left(x_i - \overline{x_w}\right)^2$$

Dies in Gleichung (10) eingesetzt:

$$s_p^{\,2} = s^2 + \frac{1}{n} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n}\left(x_i - \overline{x_w}\right)^2$$

Hierin Gleichung (3) eingesetzt:

$$s_p^{\,2} = s^2 + \frac{1}{n} \cdot s_p^{\,2}$$

$$s_p^{\,2} - \frac{s_p^{\,2}}{n} = s^2$$

$$\frac{n \cdot s_p^{\,2} - s_p^{\,2}}{n} = s^2$$

$$s_p^{\,2} \cdot \frac{n-1}{n} = s^2$$

$$s_p^{\,2} = \frac{n}{n-1} \cdot s^2$$

Hierin Gleichung (1) eingesetzt:

$$s_p^{\,2} = \frac{n}{n-1} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Damit ergibt sich als Bestimmungsgleichung für die Stichprobenvarianz:

(11) $$s_p^{\,2} = \frac{1}{n-1} \cdot \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

Die Summe der quadrierten Abweichungen wird also bei der Stichprobenvarianz nicht durch die Anzahl der erhobenen Merkmalsausprägungen *n* geteilt, sondern durch *n* – 1.

[1] Vgl. hierzu auch W. Böhme, Erscheinungsformen und Gesetze des Zufalls, Braunschweig 1964, S. 62 f.